APPLICATION OF MACHINE LEARNING FOR SPATIAL DOWNSCALING OF SATELLITE PRECIPITATION DATA (CHIRPS) OVER STATE OF MAHARASHTRA IN INDIA

By

Milind Choudhary

Under the guidance of
Dr. Neeti
Department of Natural Resources

Report Submitted
In partial fulfilment of requirements for the
Internship at Department Of Natural Resources
TERI School of Advanced Studies

teri school of
advanced studies
(Deemed to be University)

(established under Section 3 of the UGC Act. 1956)
Accredited by NAAC

TERI School of
Advanced Studies
10, Institutional Area
Vasant Kunj
New Delhi - 110 070

Tel. 7180 0222
E-mail registrar@terisas.ac.in
Fax 2612 2874
India +91 • Delhi (0) 11
Web www.terisas.ac.in

Dr. Neeti
Assistant Professor
Department of Natural Recourses
E-mail: neeti@terisas.ac.in

16th July 2019

CERTIFICATE OF INTERNSHIP

This is to certify that the report titled "APPLICATION OF MACHINE LEARNING FOR SPATIAL DOWNSCALING OF SATELLITE PRECIPITATION DATA (CHIRPS) OVER STATE OF MAHARASHTRA IN INDIA" is a record of work carried out by Milind Choudhary submitted in partial fulfilment of the requirement of Internship (May 2019- July2019) at Department of Natural Resources, TERI School of Advanced Studies. This work has been carried out under my supervision.

I wish him all success in life.

Neeti
(Neeti)

Abstract

The Climate Hazards Group Infrared Precipitation (CHIRP) is a high-resolution climatic database of precipitation obtained through satellite imaging .The difference of CHIRP database with all other existing precipitation databases is its inherent quality of high resolution which is $0.05°$. Yearly data for the period of January 2015 to December 2015 has been analysed in this study. The main aim of the study is to propose downscaling algorithms to obtain precipitation data at high resolution of 1km X 1km by establishing a relationship of CHIRP data with other variables such as vegetation, land surface temperature, elevation, latitude and longitude. The study utilised Multiple Linear Regression (MLR), Random Forest (RF) and Support Vector Machines (SVM) algorithms for downscaling. The final result has been validated with observations from rain gauge at meteorological stations. In this study it has been seen that RF has higher accuracy followed by SVM and MLR was found to be least accurate.

1. Introduction

Knowledge about precipitation data and precipitation pattern is important in the understanding of hydrological balance. It is also important for water management in agriculture, industry, power generation and drought monitoring. The dealing with extreme events and adopting appropriate mitigation measures can also be aided with proper knowledge about precipitation. Apart from this, most of the land surface processes are controlled by precipitation [1]. Thus, acquiring accurate and high resolution precipitation data is highly essential. Although, observations from meteorological stations and rain gauges are highly important in acquiring precipitation data but lack of such stations in underdeveloped areas or area with sparse rain gauge network is matter of huge concern. Application of satellites in monitoring the precipitation data has extended its reach to predict the precipitation. However, their spatial resolutions are too coarse to conclude facts at finer resolutions. This makes it difficult to formulate plans and adopt mitigation strategies at village or town level.

Various spatial downscaling models have been developed using different approaches such as linear regression model, artificial neural network [1] and so forth. Comparisons have been made among various approaches like MLR, Exponential Regression (ER), SVM and RF, by Wenlong Jing et el. They found that RF-based model was the most accurate [2]. The purpose of this study is to obtain annual cumulative precipitation maps with fine spatial resolution from coarse resolution satellite- based precipitation datasets. In this study, latitude and longitude were introduced as factors for enhancing the precipitation- land surface relationships while using any machine learning algorithms. Climatic conditions tend to remain same across same latitude with minute changes. On the other hand, latitude and longitude might affect the land surface characteristics such as soil moisture and land surface temperature [3].

Machine learning algorithms have been widely used in remote sensing image processing, land surface parameters derivation and are hence distinguished in dealing with complex and non-linear problems [4]. In this study machine learning algorithms such as Random Forest (RF), Support Vector Machine (SVM) and Multiple Linear Regression (MLR) have been applied for the downscaling process. Significance of the variables towards the models was also considered in this study.

2. Literature Review

Various studies carried out on understanding landscape suggest that precipitation data is one the most important estimators for ecological, hydrological and climatological models [2]. Furthermore, spatial and temporal distribution of precipitation is important for management of water [1].Rain Gauge and Metrological Stations have been established at various locations to know about the factual information about the precipitation. Many such rain gauge stations are present throughout the state which contains huge records of precipitation data. As mentioned by Wenlong Jing [1], problem arises in areas with sparse rain gauge network. Remote Sensing technology has been very useful for such unreachable locations where it is difficult to have rain gauge network. Remotely sensed data provide synoptic and systematic measurement over large area with least amount of human intervention at the point of data collection. However, most of the rainfall data is available coarse spatial resolution such as Tropical Rainfall Measuring Mission (TRMM) and Global Precipitation Measurement (GPCP). CHIRPS has the finest spatial resolution (5.5 km) among all the available datasets available through remote sensing technology. CHIRPS use both satellite based estimates as well as rain gauge value to provide final rainfall data. However, most of the changes in landscape such as forest conversion to agriculture or urban etc. occur at much finer scale. Therefore, change in rainfall need to be considered at fine scale compared to coarse scale. Similarly natural disasters assessment such as flood and drought assessment for which rainfall is an important component need to be carried out at fine scale. One way to make the rainfall data at the scale of landscape/local level change is through downscaling. Downscaling the available satellite data to finer resolution helps in getting data of those sparse regions [2].

Monitoring of different disasters such as flood and drought and weather prediction can also be done by analysing the pattern and distribution of precipitation [1]. Furthermore, studies have also been done in past regarding weather prediction with precipitation being an important factor [5]. Bo Pang [6], divides the statistical downscaling models into three categories: weather typing, weather generators and regression based methods [6]. Regression based methods include both linear and non-linear approaches which include MLR, Exponential Regression (ER) and other such methods. Adding machine learning algorithms to this division we can introduce non-linear and non-parametric algorithms such as SVM and RF to this classification [4]. Earlier instances of integrating latitude and longitude with other factors include the work of Shaodan Chen [3] where he uses latitude and longitude to increase the accuracy of downscaling algorithm for soil moisture. Additionally; importance of predictor selection is well explained Bo pang [6] in his research article. This coupled with proper selection of factors will add to the accuracy of the model.

3. Aims and Objectives

3.1. Comparison of various machine learning algorithms on downscaling precipitation data from 5km X 5km to 1km X 1km.

3.2. To study the impact of various variables such as Normalized Difference Vegetation Index (NDVI), Shuttle Radar Topographic Mission (SRTM), Land Surface Temperature (LST), Latitude and longitude on characteristics of rainfall.

4. Study Area

Maharashtra is a state of India situated in the western part of the country is the third largest state in the nation. The boundary has an East-West stretch of about 800km and North-South stretch of about 720km, covering a total area of 307.762 x 10^3 km^2 between 15° 44’N ~20° 60’N latitudes and 72° 36’ ~ 80° 54’ longitudes [7].

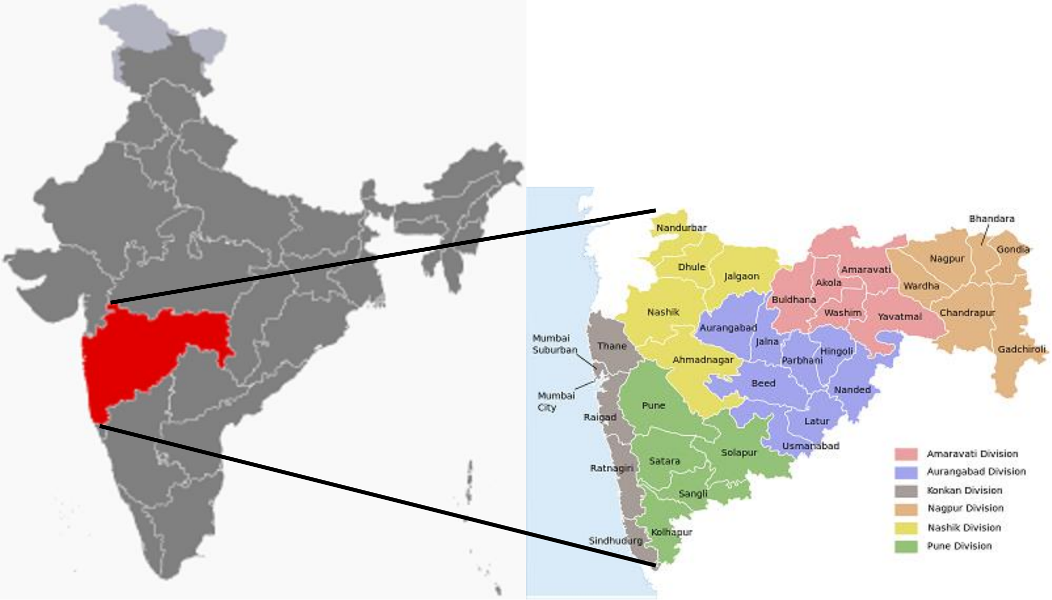


Figure 2. Map of India and Maharashtra along with divisions [8-9].

5. Data Resources

5.1. Climate Hazards Group Infrared Precipitation (CHIRP)

“The CHIRP algorithm combines three main data sources: (a) the Climate Hazards group Precipitation climatology (CHPclim), a global precipitation climatology at 0.05° latitude/longitude resolution estimated for each month based on station data, averaged satellite observations, elevation, latitude and longitude” (Funk et al., 2012; 2015b)[10]; (b) TIR-based satellite precipitation estimates (IRP); and (c) in situ rain-gauge measurements. The CHPclim is distinct from other precipitation climatologies in that it uses long-term average satellite rainfall fields as a guide to deriving climatological surfaces and was downloaded from its official website [10-11]. The product was projected into EPSG: 32643 WGS 84 / UTM zone 43N. Nearest neighbour resampling was applied to the product to bring the pixel size of the CHIRP data to 5km x 5km.

5.2. Normalised Difference Vegetation Index (NDVI)

Moderate Resolution Imaging Spectroradiometer (MODIS) product, NDVI is a vegetation index derived from atmospherically-corrected reflectance in the red, near-infrared and blue wavebands. The data was downloaded from NASA Land Processes Distributed Active Archive Center (LP DAAC) [12]. The product was projected into EPSG: 32643 WGS 84 / UTM zone 43N.

5.3. Shuttle Radar Topographic Mission(SRTM)

The DEM (Digital Elevation Model) data used in this study was obtained from NASA SRTM [13]. The SRTM has good horizontal and vertical accuracies, because of which it has been used in the study. The data obtained from SRTM was also projected into EPSG: 32643 WGS 84 / UTM zone 43N.

5.4. Land Surface Temperature (LST)

This MODIS product is an index for land surface temperature. The data was downloaded from NASA Land Processes Distributed Active Archive Center (LP DAAC)[14].The data is retrieved by the generalised split window algorithm, in which emissivities in bands 31 and 32 are estimated from land cover types, atmospheric column water vapour and lower boundary air surface temperature are separated into tractable sub-ranges for optimal retrieval. The product was projected into EPSG: 32643 WGS 84 / UTM zone 43N.

5.5. Latitude And Longitude

Latitude and Longitude were available along with all of the other satellite products [11-14]. These factors were independently extracted and used as variables for the various downscaling algorithms.

6. Methodology

6.1. Downscaling methodology

The spatial downscaling method used in the study is based on the relationship between precipitation data (CHIRP) and land characteristics. A relationship was established between precipitation and land characteristics at a coarse resolution i.e. 5km X 5km; then the established model was used to predict precipitation at a finer resolution i.e. 1km X 1km with the help of fine spatial resolution land characteristics. For downscaling the CHIRP data five land characteristics or five variables, NDVI, LST, SRTM DEM, Latitude and Longitude were used. Three machine learning algorithms were implemented, namely Multiple Linear Regression (MLR), Support Vector Machines (SVM) and Random Forest (RF), to predict the possible relationship between the precipitation data and the land characteristics. The process is described in a nutshell below:

- R programming environment along with QGIS has been used for the whole study.
- The products were converted into the same projection i.e. EPSG:32643 WGS 84 / UTM zone 43N taking help of the raster library and projectRaster function and resampling was done to get same resolution for all the products.
- The downloaded data were all converted into same extent and dimension using the process of clipping and masking using a known shapefile of Maharashtra boundary. The clipping and masking process is done using the functions crop and mask in raster package.
- The relationship between the re-sampled independent variables and CHIRP precipitation data is established by using the MLR, SVM and RF algorithms. The MLR function was pre-built into the R environment while prediction using the SVM was done by making use of e1071 library which is used for medium-scale supervised and unsupervised problems. Similarly, RF was applied by taking help from the random Forest package.
- The established models from the above mentioned machine learning algorithms was used on high spatial resolution land characteristics data i.e. 1km X 1km to achieve downscaled precipitation at 1km X 1km.
- Thereafter, the validation and error analysis was carried out using the statistical measures like Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Root Mean Squared Error (RMSE) and R^2. These statistical measures are a part of stats and Metrics package of R.

In this section, a flowchart is provided to illustrate the main steps involved in the downscaling process. It should be noted that NDVI$_{1km}$, LST$_{1km}$, SRTM$_{1km}$, and Long$_{1km}$ have been processed in the similar fashion as the products at 5km. Furthermore Table 1 contains the list of abbreviations used in the image.

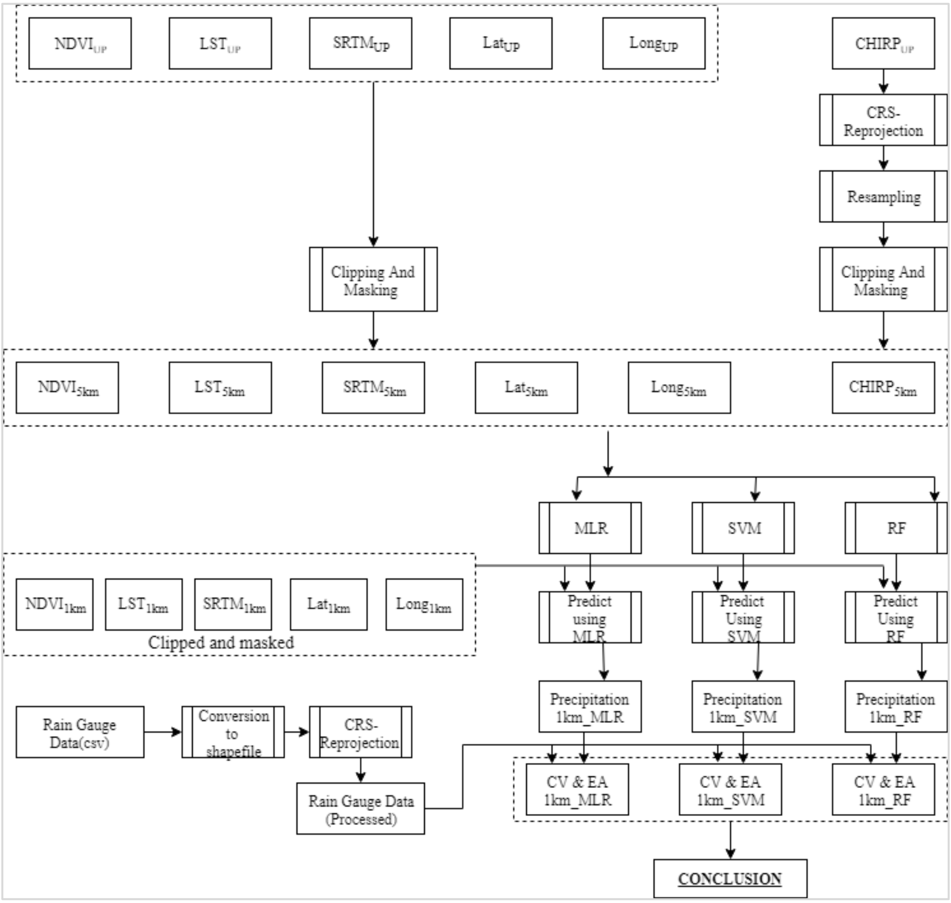


Figure 3. Flowchart of the downscaling process used in the study.

Table 1. List of Abbreviations

Acronym	Short Description
UP	Unprocessed
NDVI	Normalised Difference Vegetation Index
LST	Land Surface Temperature
SRTM	Shuttle Radar Topographic Mission(Elevation)
Lat	Latitude
Long	Longitude
CHIRP	Climate Hazards Group InfraRed Precipitation
MLR	Multiple Linear Regression
SVM	Support Vector Machine
RF	Random Forest
CV	Cross Validation
EA	Error Analysis

6.2. Multiple Linear Regression(MLR)

In this study, a MLR model was generated with NDVI, LST, SRTM, Latitude and Longitude as the variables. If P be the generated regression function and a_1, a_2, a_3, a_4 and a_5 be the slopes of each variable and c, be the intercept of the regression function then the model is:

$$\boxed{P = a_1 * NDVI_{5km} + a_2 * LST_{5km} + a_3 * SRTM_{5km} + a_4 * Latitude + a_5 * Longitude + c \qquad (1)}$$

6.3. Support Vector Machines (SVM)

SVM for regression was proposed by in 1996 by Vladimir N. Vapnik, Harris Drucker, Christopher J. C. Burges, Linda Kaufman and Alexander J. Smola[15]. The model produced by SVR depends only on a subset of the training data, because the cost function for building the model ignores any training data close to the model prediction. Apart from this the support vector machine for regression works in a similar fashion as compared to support vector machine for classification. The basic concept behind the algorithm is derived from the optimisation theory, which uses a hyperplane to classify the input variables into an m-dimensional feature space [2]. This feature space has a maximal margin in order to increase the accuracy of the model. The maximal margin is found by solving a constrained quadratic equation which is found after the applying Lagrangian methods to the decision rule of SVM along with constraints:

$$Maximize\ W(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \qquad (2)$$

Where $x_i \epsilon R_d$ are the training data and $K(x_i, x_j)$ is the kernel function.

$$f(x, \omega) = \sum_{j=1}^{m} \omega_j g_j(x) + b \qquad (3)$$

Where $g_j(x)$, j =1, 2, …., m is the set of nonlinear transformations and b is the “bias term“.

6.4. Random Forest (RF)

Random Forests as proposed by Leo Breiman is an ensemble learning method which can be used for regression by constructing a multitude of decision trees at training time and outputting mean prediction of the individual trees [16]. The forest is a combination of tree predictors such that each tree depends on randomly chosen set of subset of input variables and randomly chosen number of branches for each node. The tree predictor is based on classification and regression trees (CART) algorithm [17]. This algorithm basically constructs a tree- like graph or model of decisions and their possible consequences by recursively taking partitions of training dataset to the maximum variance between variables in the terminal nodes of the tree which are both independent and dependent. The Random Forest regression process can summarised as

- The number of trees samples set is randomly drawn from the original training set with replacement.
- Each sample set is a bootstrap sample and the samples not included in the bootstrap are said as out-of-bag data (OOB).
- For each bootstrap sample a regression tree is grown with the modification that at each node random subset of variables is selected from which the best variable is chosen for the node.
- Prediction for new samples is made by averaging the predictions from each of the trees:

$$f = \frac{1}{N}\sum_{i=1}^{N} f_i(x) \qquad (4)$$

Where N is the number of trees and $f_i(x)$ is the prediction from each regression tree.

6.5. Validation

Validation was done on the basis of 22 metrological stations distributed in the study area as shown in Figure 1. Three comparison criteria were, the coefficient of determination (R^2), the mean absolute error (MAE), and the root mean squared error (RMSE), which are expressed as

$$R^2 = \frac{\{\sum_{k=1}^{n}[(Y_k - \bar{Y})(O_k - \bar{O})]\}}{\sqrt{([\sum_{k=1}^{n}(Y_k - \bar{Y})^2])([\sum_{k=1}^{n}(O_k - \bar{O})^2])}} \qquad (5)$$

$$MAPE = \frac{\sum_{k=1}^{n}|(Y_k - O_k)|}{n} \qquad (6)$$

$$RMSE = \sqrt{\frac{\sum_{k=1}^{n}(Y_k - O_k)^2}{n}} \qquad (7)$$

Where, Y_k is the observation measured by the station k,

O_k is the precipitation predicted by a model at location of station k,

$\bar{Y}$ Is the mean value of all the observations recorded at the stations,

And $\bar{O}$ is the mean value of the predicted precipitation at all station locations [2].

7. Data Analysis And Results

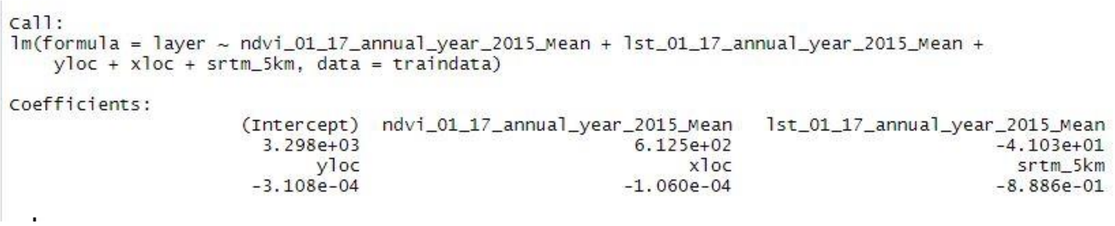
The CHIRPS data was downscaled from 5km to 1km using the algorithm based on MLR, SVM and RF. The MLR and SVM were performed using a combination of all the variables that are NDVI, SRTM, LST, Latitude and Longitude. RF was performed using two approaches that are:

i. NDVI, SRTM, LST, Latitude and Longitude as variables.
ii. NDVI, SRTM, LST, and Longitude as variables.

It should be noted that the abbreviations used for variables and their corresponding terms in the study is given in Table 3.

7.1. Multiple Linear Regression

The values of coefficients in Equation (1) for the multiple linear regression model came out to be $a_1 = 6.125 \text{ X } 10^{2}$, $a_2 = -4.103 \text{ X } 10^{1}$, $a_3 = -8.886 \text{ X } 10^{-1}$, $a_4 = -1.060 \text{ X } 10^{-4}$, $a_5 = -3.108 \text{ X } 10^{-4}$ and $c = 3.298 \text{ X } 10^{3}$ (Figure 4)(Table 2). MLR gave the least accurate results when compared to the other three models. Hence, it was used as a test model to compare accuracies of other models.



```
Call:
lm(formula = layer ~ ndvi_01_17_annual_year_2015_Mean + lst_01_17_annual_year_2015_Mean +
    yloc + xloc + srtm_5km, data = traindata)

Coefficients:
                 (Intercept)  ndvi_01_17_annual_year_2015_Mean   lst_01_17_annual_year_2015_Mean
                   3.298e+03                         6.125e+02                        -4.103e+01
                        yloc                              xloc                          srtm_5km
                  -3.108e-04                        -1.060e-04                        -8.886e-01
```

.

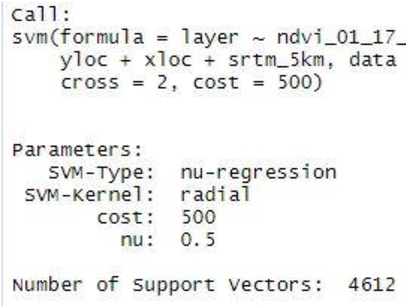
Table 2. Coefficient Of MLR and values

Variable	Coefficient
a_1 (NDVI)	$6.125 \text{ X } 10^{2}$
a_2 (LST)	$-4.103 \text{ X } 10^{1}$
a_3 (SRTM DEM)	$-8.886 \text{ X } 10^{-1}$
a_4 (Latitude)	$-1.060 \text{ X } 10^{-4}$
a_5 (Longitude)	$-3.108 \text{ X } 10^{-4}$
c (Intercept)	$3.298 \text{ X } 10^{3}$

Figure 4. Details of the linear model.

7.2. Support Vector Machine

Table 4 gives the parameter combination used for SVM. The best fit SVM model was of type nu-regression, kernel was of radial basis type with cost as 500 and nu as 0.5(Figure 5). SVM-based model increased the accuracy as compared to MLR but still performed inferior to the two models developed using random forest approach.



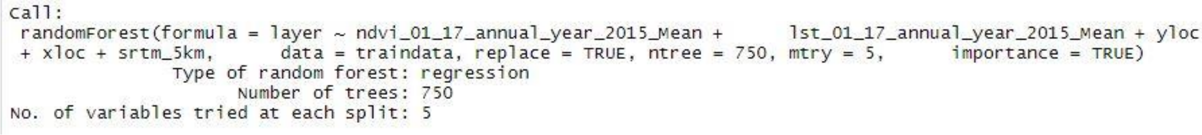
```
Call:
svm(formula = layer ~ ndvi_01_17_annual_year_2015_Mean + lst_01_17_annual_year_2015_Mean +
    yloc + xloc + srtm_5km, data = traindata, type = "nu-regression", kernel = "radial",
    cross = 2, cost = 500)


Parameters:
   SVM-Type:  nu-regression
 SVM-Kernel:  radial
       cost:  500
         nu:  0.5

Number of Support Vectors:  4612
```

Figure 5. Details of the SVM model.

7.3. Random Forest

7.3.1. RF(5 variables)

The first random forest model developed used all the variables and the parameter combination tested is given in Table 4. The best fit had 750 trees and 5 variables tried at each split (Figure 6). This model was the second best model developed in the study.

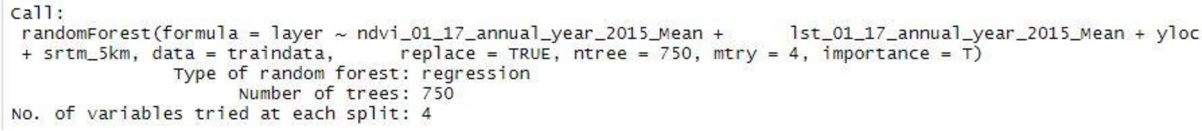


```
Call:
 randomForest(formula = layer ~ ndvi_01_17_annual_year_2015_Mean +      lst_01_17_annual_year_2015_Mean + yloc
 + xloc + srtm_5km,      data = traindata, replace = TRUE, ntree = 750, mtry = 5,      importance = TRUE)
               Type of random forest: regression
                     Number of trees: 750
No. of variables tried at each split: 5
```

Figure 6. Details of RF(5 variables)

7.3.2. RF(4 variables)

The second and most accurate model of the study was developed using four variables: NDVI, LST, SRTM DEM and Longitude. The parameter combination for the same is given in Table 4. The best fit model had 750 trees and 4 variables tested at each split (Figure 7). This model gave the most similar results at finer resolutions as compared to the CHIRP dataset which was at a coarser resolution.



```
Call:
 randomForest(formula = layer ~ ndvi_01_17_annual_year_2015_Mean +      lst_01_17_annual_year_2015_Mean + yloc
 + srtm_5km, data = traindata,      replace = TRUE, ntree = 750, mtry = 4, importance = T)
               Type of random forest: regression
                     Number of trees: 750
No. of variables tried at each split: 4
```

Figure 7. Details of RF(4 variables).

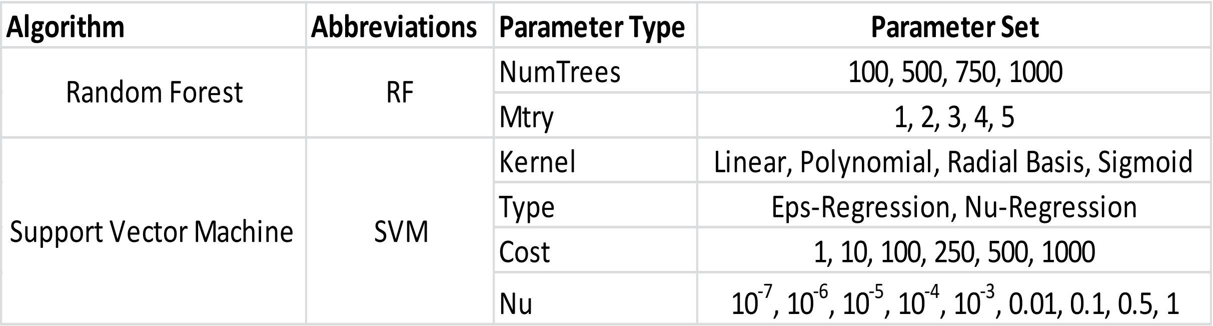
Table 4. List of Variable names used in models and corresponding names in study.

Variable Name in Model	Name used in Study
ndvi_01_17_annual_year_2015_Mean	NDVI
lst_01_17_annual_year_2015_Mean	LST
Srtm_5km	SRTM DEM
Xloc	Latitude
Yloc	Longitude

7.4. Comparative Study of Results

The RF and SVM models depend significantly on the choice of various arguments for their respective functions. The choice of optimal parameters and number of variables were highly essential. Tuning process was done for combinations of parameters using tune function in the e1071 package of R and the process returned an optimal set of parameters using which we predicted at a resolution of 1km X 1km. On the other hand it should be noted that all the variables were used as independent variables for the MLR model. Table 4 contains a list of the values for the parameters of RF and SVM model. The explanation of parameters is given in table 5.

Table 4. Parameter Combination for each algorithm.



Algorithm	Abbreviations	Parameter Type	Parameter Set
Random Forest	RF	NumTrees	100, 500, 750, 1000
		Mtry	1, 2, 3, 4, 5
Support Vector Machine	SVM	Kernel	Linear, Polynomial, Radial Basis, Sigmoid
		Type	Eps-Regression, Nu-Regression
		Cost	1, 10, 100, 250, 500, 1000
		Nu	10^{-7}, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 0.01, 0.1, 0.5, 1

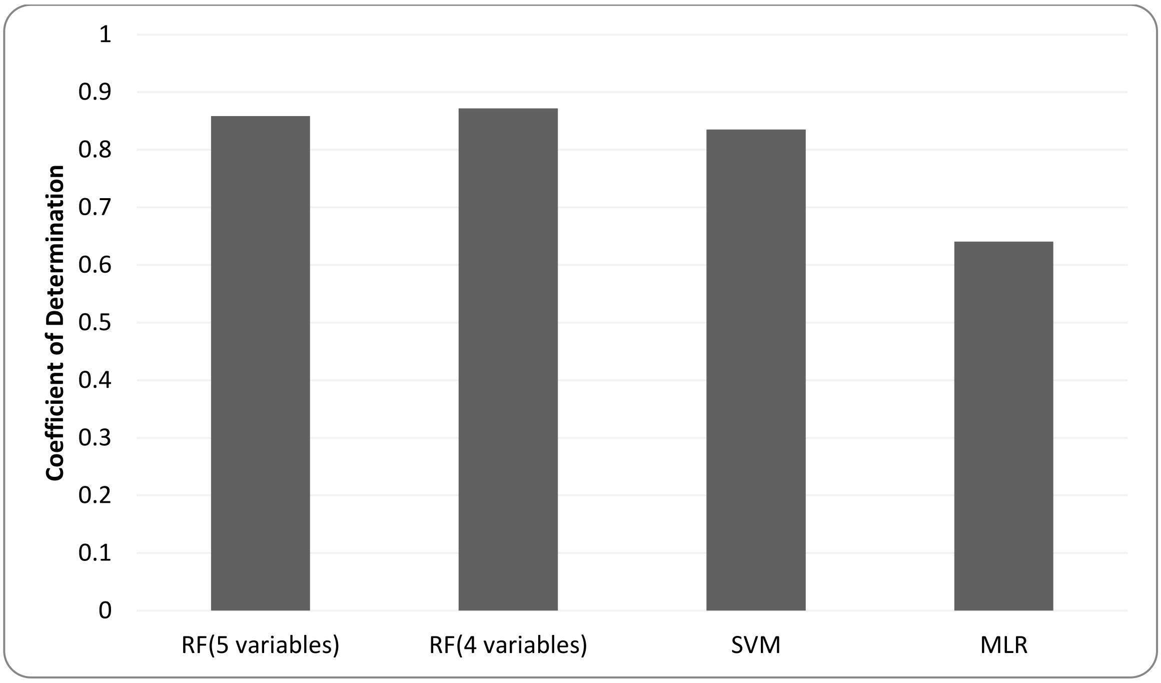
Table 4. Explanation of Parameters in Table 3.

Parameter Type	Meaning of Parameter
NumTrees	Number of trees to be considered while developing the random forest.
Mtry	Number of variables to be considered at each split.
Kernel	Kernel used in SVM for construction of hyperplane.
Type	Set the SVM function for regression or classification based on Eps and Nu type.
Cost	Cost of constraints violation. It is the 'C'-constant of the

	regularization term in the Lagrange formulation.
Nu	Parameter needed for nu regression. Its value ∈ (0,1).

Figure 8 presents the R^2, MAE and RMSE estimated by model for the year 2015. In general RF- based model with NDVI, SRTM, LST and longitude as variables (RF(5 variables)) produced the highest R^2 and the lowest MAE and RMSE, followed by the RF-based model with NDVI, SRTM, LST, Latitude and Longitude as variables (RF(4 variables)). This indicated that exclusion of Latitude is beneficial for increasing model accuracy. Nu- regression SVM –based model was third best in terms of accuracy followed by MLR which produced the least accurate model.

(a)



(a)

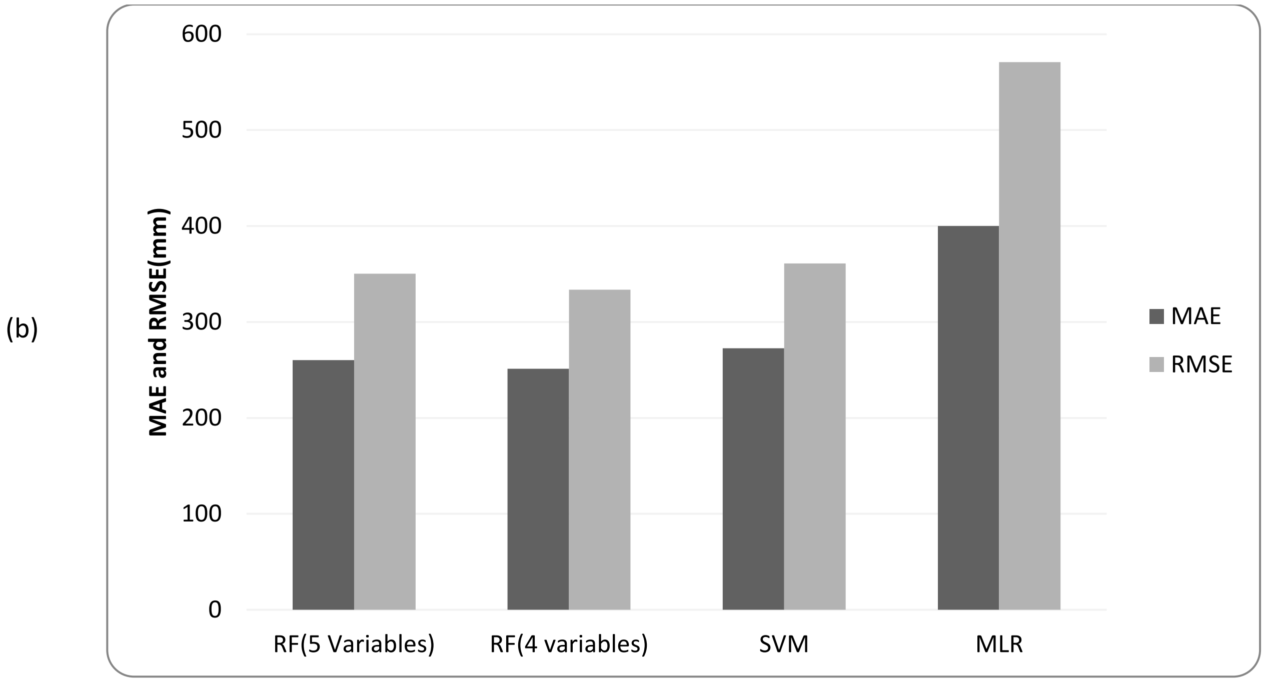
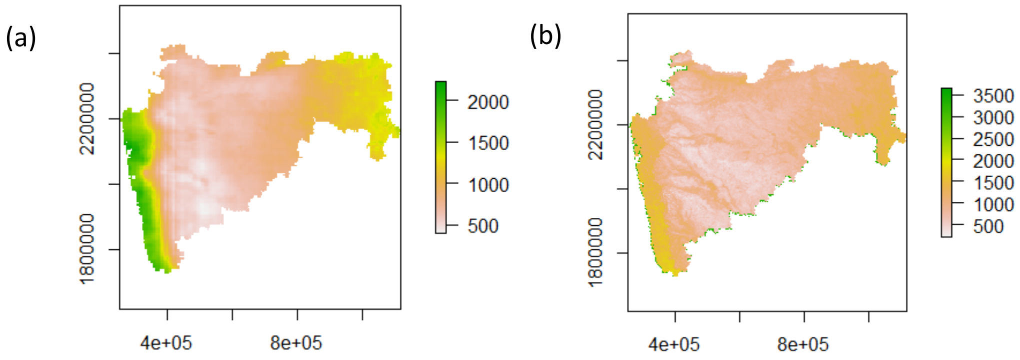


Figure 8. (a) Coefficient of Determination (R^2), (b) Mean Absolute Error (MAE) and Root Mean Square Error(RMSE) predicted values using different algorithms compared with original rain gauge data.

Figure 9 shows the CHIRP data of the Maharashtra region and the downscaled results using the MLR, SVM, RF(5 variables) and RF(4 variables). SVM, RF(5 variables) and RF(4 variables) have spatial distribution patterns similar to those of CHIRP which indicates their accuracy in prediction of precipitation data.



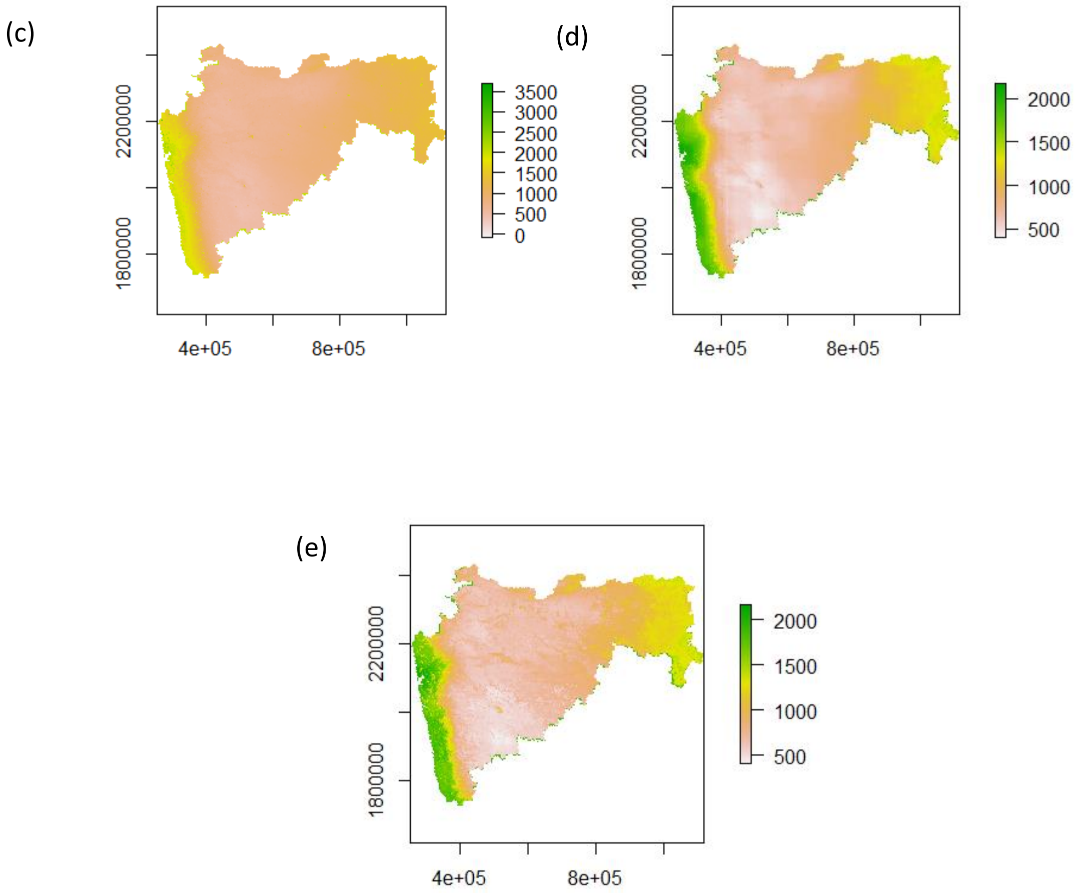


Figure 9. (a)CHIRP precipitation data and downscaled results of (b)MLR, (c) SVM, (d) RF(5 variables), (e) RF(4 variables)

7.5. Analysis Of Error and Validation

7.5.1. Validation with Rain Gauge Observations

The downscaled results of each algorithm were validated with rain gauge data from 22 sites across the state of Maharashtra. Figure 10 contains the scatter plot of CHIRP data along with downscaled results of each algorithm. The scatter plot of RF(4 variables) was very similar to the scatter plot of the CHIRP. This implies that the performance of RF(4 variables) was better compared to other models.

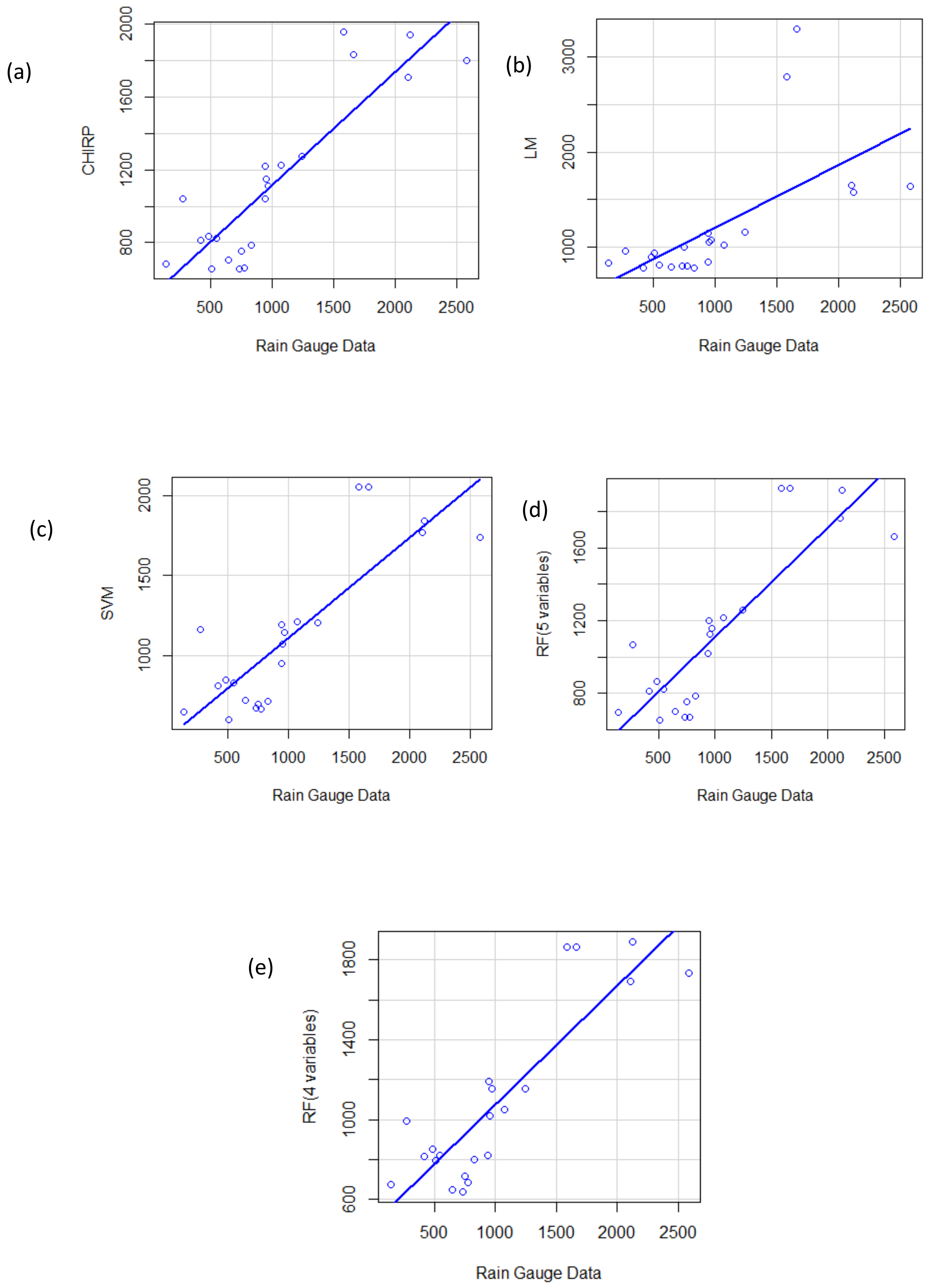
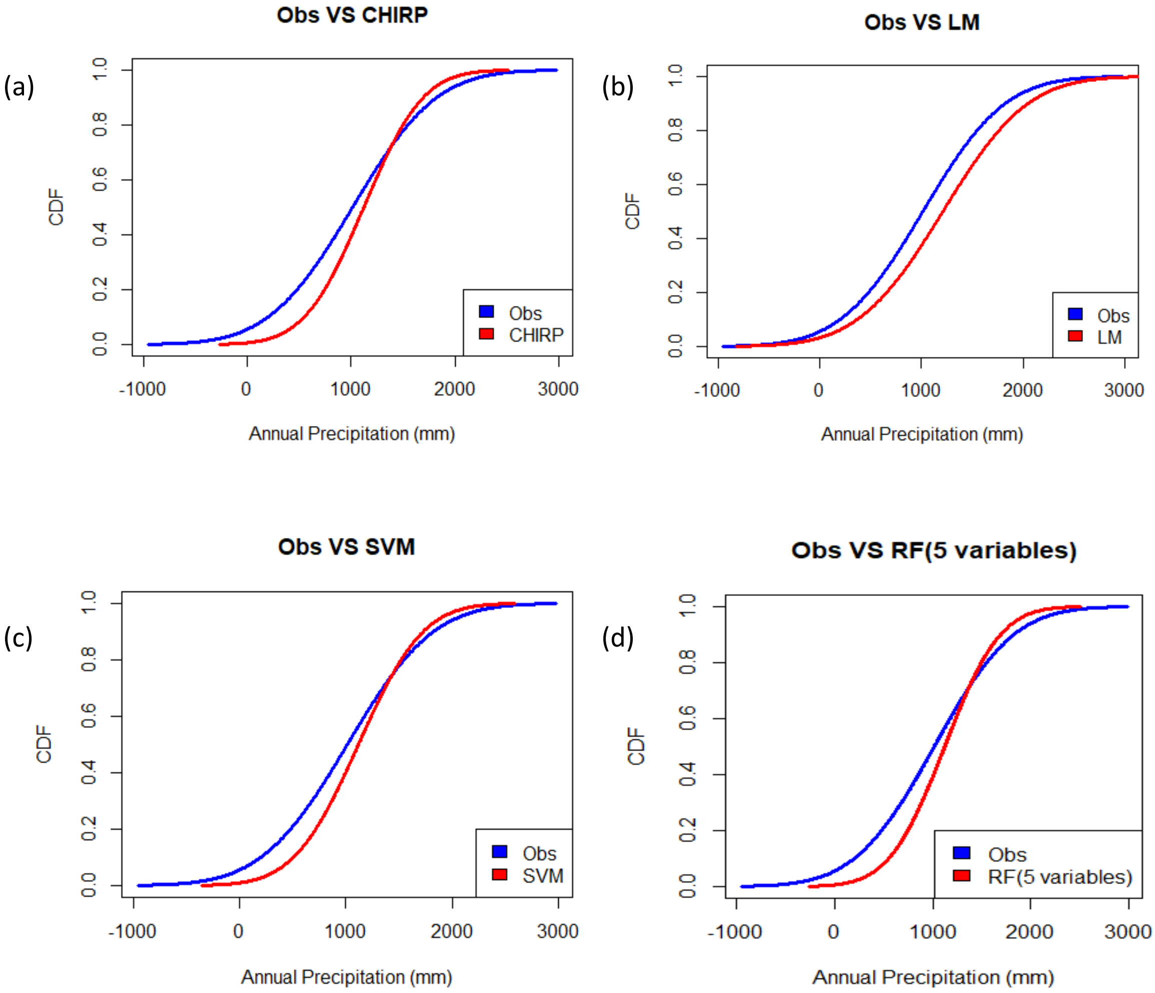


Figure 10. Scatter Plot of (a)CHIRP precipitation data and downscaled results of (b)MLR, (c) SVM, (d) RF(5 variables), (e) RF(4 variables)

Figure 11a-e shows the Cumulative Distributive Function (CDF) of observations measured by stations with CHIRP and downscaled results from each model. In general, if the shape of CDF of the Observed (Obs) and the predicted values is similar to the shape of Obs and CHIRP, this implies that the respective model has better accuracy for prediction of precipitation data. The original CHIRP was estimating the precipitation over the state of Maharashtra with R^2= 0.8801156, MAE= 253.0468mm and RMSE = 332.0902mm. The R^2 of RF(4 variables) was the closest to the CHIRP precipitation data with a value of 0.871646. The MAE and RMSE of the same were 251.4124mm and 333.816mm respectively. RF(5 variables) was the second most accurate model with R^2= 0.8582818, MAE= 260.2705mm and RMSE= 350.1675mm. SVM followed next with R^2= 0.835408, MAE= 272.7378mm and RMSE= 360.9244mm.Finally, the least accurate model was MLR with R^2= 0.6407607, MAE= 400.0735mm and RMSE= 570.978mm.



(e)

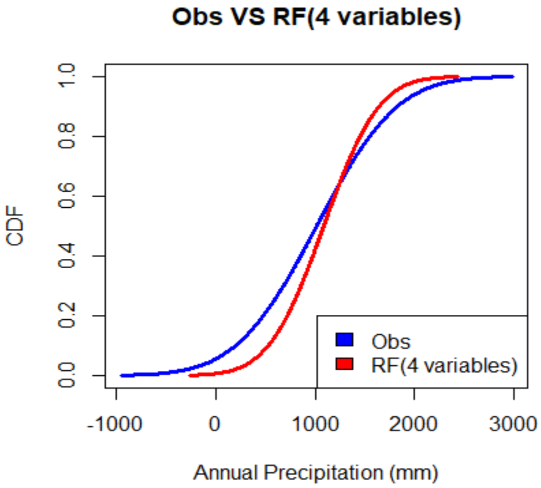


Figure 11. CDF of the observations and (a)CHIRP precipitation data and downscaled results of (b)MLR, (c) SVM, (d) RF(5 variables), (e) RF(4 variables)

7.5.2. Variable Importance of Random Forest Model

The RF model gives the argument in R to measure the variable importance. A graph was plotted which gave us %IncMSE and IncNodePurity. According to trend more the values of these indicators more is the importance of variable in the model. In contrast to this trend, removal of xloc or latitude resulted in significant improvement of the model. The error and validation process was done for models developed by removal of variables lie NDVI, SRTM and combination of variables but such accuracy was not achieved. So the model with all the variables (RF(5 variables)) and model with all but latitude (RF(4 variables)) were considered for the development of models and prediction of the precipitation data at a resolution of 1km X 1km.

Figure 12(a) shows the variable importance plot with all the factors and gives the result that xloc (latitude) is the most important variable. On the other hand, 12(b) shows the variable importance plot after removal xloc or latitude and gives the result that SRTM DEM and LST are the most important variables. Table 3 contains the list of variable name used in development of model and their corresponding names used in the study.

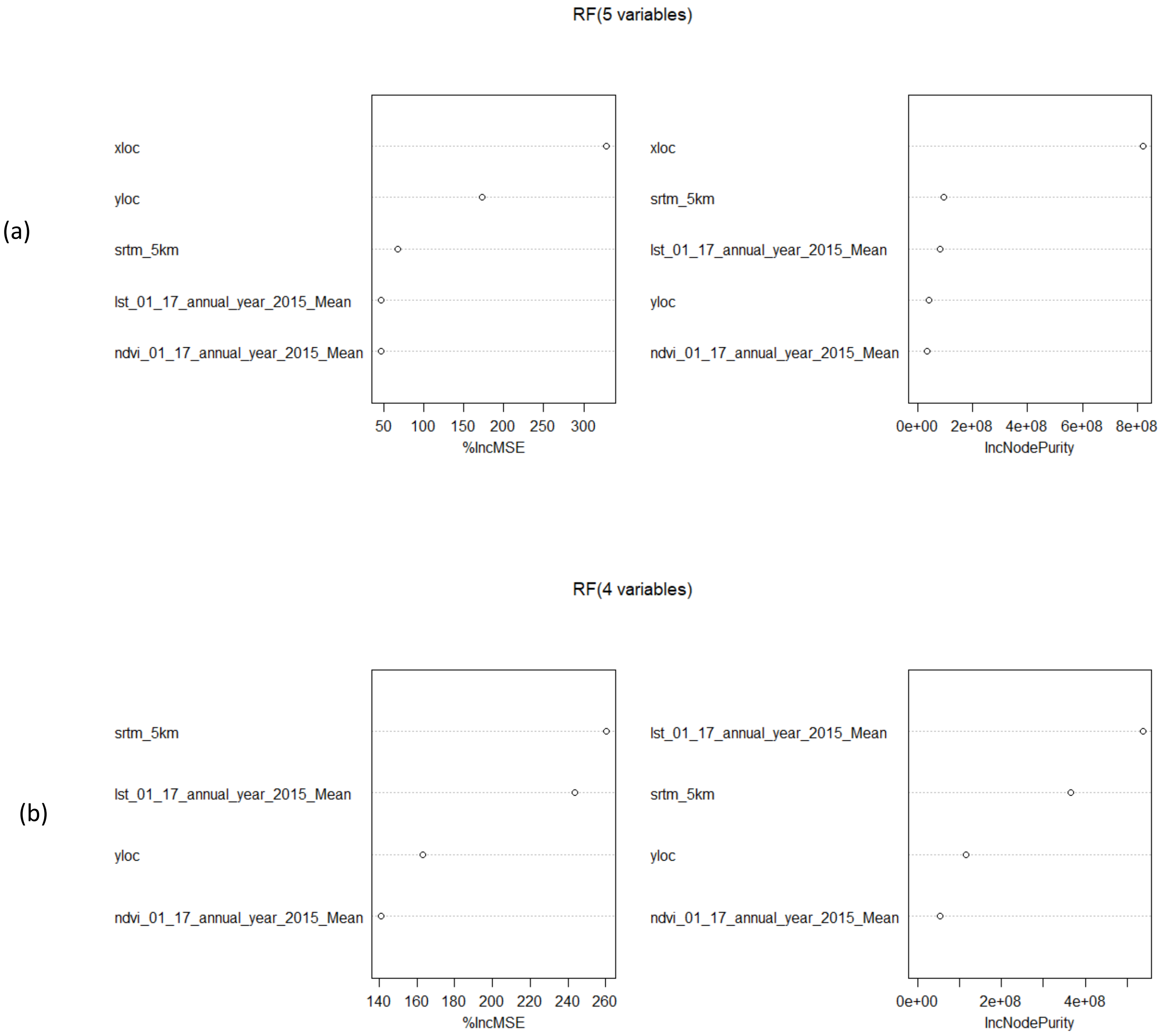


Figure 12.(a) Variable Importance Plot for RF(5 variables). (b)Variable importance plot for RF(4 variables).

8. Discussion

8.1. Value of Spatial Downscaling

Precipitation is the most active flux and greatest input to near surface hydrological system and thus strongly influences hydrological states and fluxes. Quantification of the spatial distribution of precipitation is thus significant to quantify these states and fluxes. Good estimates of the spatial variability of precipitation are especially crucial for accurate prediction of runoff response [18]. However, precipitation data at a resolution of 1kmX1km is not available for most of the land extent especially in places where the rain gauge stations are sparsely located. Satellite data can be used for the purpose of getting precipitation data for the whole land terrain but it is often limited by the coarse spatial resolution which in this study was 5kmX5km. In this study the annual rain was downscaled from 5kmX5km to 1kmX1km using machine learning algorithms like MLR, RF and SVM over the state of Maharashtra. RF model produced accuracy of similar magnitude as compared to CHIRPS data when cross validated with observations from rain gauge throughout Maharashtra.

8.2. Usability of NDVI, SRTM, LST, Latitude and Longitude for Downscaling Precipitation dataset

Vegetation of an area affects the precipitation pattern is an accepted fact [19]. Furthermore, vegetation exerts strong influence on latent heat which affects the humidity and further influencing development of atmospheric circulations on both small and large scale [20]. Thus, NDVI is commonly used as a variable for downscaling the precipitation data. However, NDVI might not be useful in areas where the land is barren with negligible to no vegetation.

The DEM which is synonymous to topography of the region is yet another factor which has been seen to affect the precipitation pattern of a region [21]. Theoretically, increase in elevation could increase the relative humidity of the air masses through expansion and cooling of the rising air masses, resulting in precipitation [21]. Therefore, SRTM is used as a variable for the downscaling models developed and used in this study.

LST has been used as another factor in the study. It has been observed that if the ground is wet; more energy is likely to be used to evaporate it [23]. Moreover, if the ground is wet from precipitation, the associated clouds likely block the sun, initially providing less energy and further reducing the temperature. In addition, high rates of evaporation could occur directly from bare soil after periods of rain, further suppressing sensible heat and surface temperature [24]. Thus, LST can be robustly used in the downscaling models.

In this study effect of latitude and longitude was also taken into consideration while developing the model. It has often been seen in theory that climatic conditions remain more or less similar in particular latitude. In similar fashion in a particular longitude there is no such established trend. But, these factors coupled with above mentioned might be a deciding factor which has been explored in the study.

8.3. Variable Selection of Models

During the development of various models it was seen that not all variables contributed to the accuracy of models as compared to rest of the factors. In development of RF model it was seen that latitude did not amount to the improvement of R^2 or the decrease of MAE and RMSE. This was against the normally seen trends since the %IncMSE and IncNodePurity of this factor was the highest.

9. Conclusion

In this study three machine learning algorithms, Random Forest (RF), Support Vector Machine (SVM) and Multiple Linear Regression (MLR) were used to downscale the yearly CHIRP data from 5km X 5km to 1km X 1km over the state of Maharashtra. Furthermore latitude and longitude were added as new variables in addition to vegetation, land surface temperature and elevation. The downscaled results were validated with observations of rain gauge from meteorological stations across the Maharashtra region.

The validation results showed that RF and SVM- based models produced higher accuracy compared to MLR model. Furthermore, RF(4 variables) followed by RF(5 variables) showed better performance compared to SVM based model. RMSE calculations for each of the model also suggest that RF(4 variables) had the highest accuracy with RMSE of 333.816mm followed by RF(5 variables) with 350.1675mm. SVM performed slightly poor as compared to RF(5 variables) with RMSE of 360.9244mm and the worst accuracy was given by MLR with RMSE 570.978mm. According to variable importance measurements in the RF models, latitude was the most significant factor, followed by SRTM and LST.

10.Scope of further study

In the future, variable importance could also be tested in the SVM- based model in order to increase the accuracy of this model. New factors related to land surface such as slope, nearness to sea, soil moistures and aspects could be incorporated into the models for downscaling the satellite data. Moreover, further study can be undertaken to see the accuracy of the models at daily or weekly or monthly basis. This will hold great importance in environmental and ecological research.

References

1. Adrianos Retalis, Filippos Tymvios, Dimitrios Katsanos, Silas Michaelides; Downscaling CHIRPS precipitation data: an artificial neural network modelling approach
2. Wenlong Jing, Yaping Yang, Xiafang Yue and Xiaodan Zhao. A Spatial Downscaling Algorithm for Satellite-Based Precipitation over the Tibetan Plateau Based on NDVI, DEM, and Land Surface Temperature.
3. Shaodan Chen, Dunxian She , Liping Zhang , Mengyao Guo and Xin Liu; Spatial Downscaling Methods of Soil Moisture Based on Multisource Remote Sensing Data and its Application.
4. Shao, Y.; Lunetta, R.S. Comparison of support vector machine, neural network, and cart algorithms for the land-cover classification using limited training data points. ISPRS J. Photogramm. Remote Sens. 2012
5. Robert J. Kuligowski and Ana P. Barros; Localized Precipitation Forecasts from a Numerical Weather Prediction Model Using Artificial Neural Networks.
6. Bo Pang, Jiajia Yue, Gang Zhao and Zongxue Xu; Statistical Downscaling of Temperature With the Random Forest Model.
7. Salunkhe P. Y. , Gharpure V. T. Ginger Cultivation in Maharashtra A Geographical Analysis.
8. Filpro [CC BY-SA 4.0 (https://creativecommons.org/licenses/by-sa/4.0)]
9. Maharashtra_locator_map.svg: User:PlaneMadderivative work: Kaajawa [CC BY-SA 3.0 (https://creativecommons.org/licenses/by-sa/3.0)]
10. Tufa Dinku , Chris Funk, Pete Peterson, Ross Maidment, Tsegaye Tadesse, Hussein Gadain, Pietro Ceccato. Validation of the CHIRPS satellite rainfall estimates over eastern Africa.
11. http://chg.geog.ucsb.edu/data/chirps/
12. https://modis.gsfc.nasa.gov/data/dataprod/mod13.php
13. http://srtm.csi.cgiar.org/srtmdata/
14. https://modis.gsfc.nasa.gov/data/dataprod/mod11.php
15. Drucker, Harris; Burges, Christopher J. C.; Kaufman, Linda; Smola, Alexander J.; and Vapnik, Vladimir N. (1997); "Support Vector Regression Machines", in Advances in Neural Information Processing Systems 9, NIPS 1996, 155–161, MIT Press.
16. Breiman, L. Random forests. Mach. Learn. 2001, 45, 5–32.

17. Breiman, L.; Friedman, J.H.; Olshen, R.A.; Stone, C.J. Classification and Regression Trees; Chapman and Hall/CRC: Boca Raton, FL, USA, 1984.
18. Guan, H.; Wilson, J.L.; Xie, H. A cluster-optimizing regression-based approach for precipitation spatial downscaling in Mountainous Terrain. J. Hydrol. 2009
19. Zhang, X.; Friedl, M.A.; Schaaf, C.B.; Strahler, A.H.; Liu, Z. Monitoring the response of vegetation phenology to precipitation in africa by coupling modis and trmm instruments. J. Geophys. Res. Atmos. 2005
20. Spracklen, D.V.; Arnold, S.R.; Taylor, C.M. Observations of increased tropical rainfall preceded by air passage over forests. Nature 2012
21. Guan, H.; Wilson, J.L.; Xie, H. A cluster-optimizing regression-based approach for precipitation spatial downscaling in Mountainous Terrain. J. Hydrol. 2009
22. Sokol, Z.; Bližňák, V. Areal distribution and precipitation-altitude relationship of heavy short-term precipitation in the Czech Republic in the warm part of the year. Atmos. Res. 2009

23. Trenberth, K.E.; Shea, D.J. Relationships between precipitation and surface temperature. Geophys. Res. Lett. 2005
24. De Kauwe, M.G.; Taylor, C.M.; Harris, P.P.; Weedon, G.P.; Ellis, R.J. Quantifying land surface temperature variability for two sahelian mesoscale regions during the wet season. J. Hydrometeorol. 2013